ETH Zurich
Institute of Neuroinformatics

Master's Thesis Opportunity
Mechanistic Interpretability of Multimodal (Audio) Language Models

Supervisors

Prof. Dr. Benjamin Grewe	ETH Zürich	Primary PI	
Roman Boehringer	ETH Zürich	Supervisor	roman@ini.ethz.ch
Niclas Pokel	ETH Zürich	Supervisor	npokel@ethz.ch
Dr. Yingqiang Gao	SDSC (ETHZ)	Supervisor	yingqiang.gao@sdsc.ethz.ch
Dr. Pehuen Moure	Cornell University	Supervisor	pehuen@cornell.edu

Abstract

Audio-language models promise to adapt to impaired speech at inference time when given clinical context about the speaker. Our previous work shows the opposite: across ten models on the Speech Accessibility Project, clinical context did not help and often degraded word error rate (WER) [1]. A follow-up mechanistic study on Gemma-4-E2B localized a repairable part of the failure to the final-input-token residual stream and introduced SCRA, a one-pass edit that closes about 60% of the gap between audio-only and clinical-context prompting [2]. That work leaves open whether clinical information is ever *used*, since SCRA only reproduces the audio-only state, and several of its comparisons are confounded. This thesis will make the mechanistic evidence rigorous, separate generic from clinical-specific prompt effects, trace the failure to its causal origin, and benchmark repairs that aim to exploit clinical context rather than merely avoid its harm.

Research Goals and Problem Statement

The goal of this master's thesis is to determine why frozen audio-language models fail to benefit from clinical context on impaired speech, and whether this can be turned into a usable gain. It addresses three problems:

1. **Validity.** The prior mechanistic study has open confounds: a patching baseline drift of about 0.009 WER (a fifth of the rescue effect), unpaired confidence intervals, a probe-versus-complement test that is not size-matched (20 vs. 1,516 coordinates), only four middle layers, and a "Difficult Samples" gain defined by conditioning on the baseline's own errors.

2. **Use of clinical information.** Wrong-content prompts perform like correct ones, and SCRA targets the audio-only state, so the method mimics ignoring the context. It is unknown whether a clinical-specific signal exists that a frozen model could exploit.

3. **Mechanism and repair.** The pathway from clinical block to degraded transcript is unexplained, and SCRA (WER 0.354) is still worse than omitting the context (0.327).

This thesis will be partially supervised by a collaborator at Cornell University. There will also be a master's thesis conducted at Cornell, which our team at ETH INI will co-supervise reciprocally. Ideally, both teams work in close collaboration with regular meetings and status reports. As a working split (to be confirmed), the ETH thesis covers diagnosis and mechanism, and the Cornell thesis covers repair and training-based adaptation.

Proposed Work, Expected Outcomes, and Results

The concrete research questions and experimental design will be shaped by the student in close discussion with both teams. Possible directions include:

- **Consolidating the mechanistic evidence:** Re-examine the prior findings with stronger controls and statistics, and test how far they generalize beyond a single model and prompt template.
- **Understanding what the context does:** Investigate whether the prompt-induced effect is generic or clinical-specific, and whether a frozen model can use clinical information at all.
- **Tracing the failure mechanism:** Identify where inside the model the degradation arises and how it propagates to the transcript.
- **Developing and benchmarking repairs:** Design inference-time or lightweight-training interventions that go beyond avoiding the harm, and compare them against simple baselines and fine-tuning, including when to apply context.

Expected Outcome: A reproducible mechanistic analysis that confirms or corrects the prior findings, with released code. A clearer answer to whether and how audio-language models can exploit clinical context, and ideally an improved repair method. A master's thesis and, ideally, a conference paper co-authored with the Cornell team. Negative findings are also valuable, as they would strengthen the case that targeted training is necessary for accessibility.

Discussion of Prior Work

This project is directly built upon the successful foundation of a previous master's thesis, which resulted in a workshop submission [2]. This thesis is meant to extend that work. The original line of work was motivated by a systematic evaluation in which we tested whether providing open multimodal LLMs with additional clinical context can improve impaired speech recognition, and therefore out-of-distribution in-context learning performance. This was published at COLM 2026 [1]: the models actually degraded even under theoretically informative context, and the degradation tracked the presence of non-instructional text rather than its content. The follow-up study localized the repairable part of this failure in Gemma-4-E2B and introduced SCRA, but covers a single model and prompt template, contains the confounds listed above, and stops at the residual stream.

Qualifications

This project is ideal for a Master's student who is highly independent, proactive, and capable of taking full ownership of their research. A strong interest in pushing the boundaries of current research and publishing the results is essential.

- Excellent programming skills in Python and deep familiarity with PyTorch.
- Strong theoretical understanding of deep learning and machine learning principles. Knowledge in natural language processing (NLP), computational neuroscience and interpretability is a big plus.
- Self-driven and motivated to work independently on a challenging research topic.
- A clear and demonstrated interest in producing a scientific publication for an international conference.

References

[1] Pehuen Moure, Niclas Pokel, Bilal Bounajma, Yingqiang Gao, Roman Boehringer, Longbiao Cheng, and Shih-Chii Liu. When audio-language models fail to leverage multimodal context for dysarthric speech recognition. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=8bdJNPLIKV.

[2] Niclas Pokel, Pehuen Moure, Bilal Bounajma, Longbiao Cheng, Yingqiang Gao, Shih-Chii Liu, and Roman Boehringer. Localizing and repairing context-failures in audio-language models. In *Actionable Interpretability Workshop at COLM 2026*, 2026. URL https://openreview.net/forum?id=daxfgcyXEk.